Us Versus Them: Ensuring Practical and Psychological Utility of Measurements of Schematic Map Usability

Maxwell. J. Roberts
Department of Psychology
University of Essex
Colchester, UK
mjr@essex.ac.uk

***Abstract*—Measures of schematic map usability should have practical utility – i.e. connection with real-world issues – and psychological utility – i.e. connection with aspects of usability that users consider to be important. The measure of *journey planning time* generally has a zero correlation with various subjective measures such as map choice. Although this *usability gap* can be explained as a metacognitive deficit, it is suggested that we may be measuring aspects of usability that have low practical and psychological utility, and that an attempt to identify other measures in order to bridge the gap may lead to improved user-acceptance of *both* new map designs *and* research findings.**

Keywords—schematic mapping; effective design; usability testing; usability gap; metacognition.

I. OBJECTIVE MEASURES OF USABILITY

Whenever the difficulty of a task is to be determined objectively, there are two obvious ways in which performance can be measured; the length of time necessary to perform the task, and the numbers of errors made while undertaking it. Pick up any experimental psychology textbook and these measures dominate in descriptions of research. The underlying logic is that there are many aspects of task difficulty that can lead to extended solution times, for example, if key elements are difficult to identify, or else many inference steps are needed in order to generate a solution. In tandem, these aspects can lead to errors when, for example, a task is incompletely understood, or inappropriate aspects of it are incorporated into the solution process, or else there is a failure at some point within a long sequence of inference steps. Hence, if we compare two tasks in which the same information is configured in different ways, and one configuration requires longer times for successful solution than the other and/or it is associated with more errors, then it can be concluded that this particular configuration is harder to utilise than the other, and is less appropriate to administer to a user in a real-world setting.

In the specific context of schematic maps, the logic translates straightforwardly[1]. Errors (invalid/impossible routes), for example, caused by misinterpreting transfer opportunities between lines, or failure to follow routes correctly, indicate configurations that are difficult to interpret. This measure of difficulty has *practical utility*: it has a clear direct link to undesirable real-world consequences, as capitalized on in one of the earliest studies of subway map usability [4]. The measure also has *psychological utility*: mistakes are undesirable to a user and hence the disadvantage of a design that is poor in this respect is easy to communicate. As another example, Lloyd, Rodgers and Roberts [11] identified a number of line configurations on the New York City Weekender subway map that potentially confronted the user with *navigational hazards,* and showed that coding lines with a common colour reflecting their *trunk route* through Manhattan led to more errors when trying to track lines than using *individual route* colours (see Fig. 1). In the context of this task, the prevalence of errors indicated that trunk route colours were more difficult for the user than individual route colours, and hence the map configured in this way was harder to use – a *less effective design*.

In general, it would be legitimate to query the fitness-for-purpose of any schematic map whose use was associated with

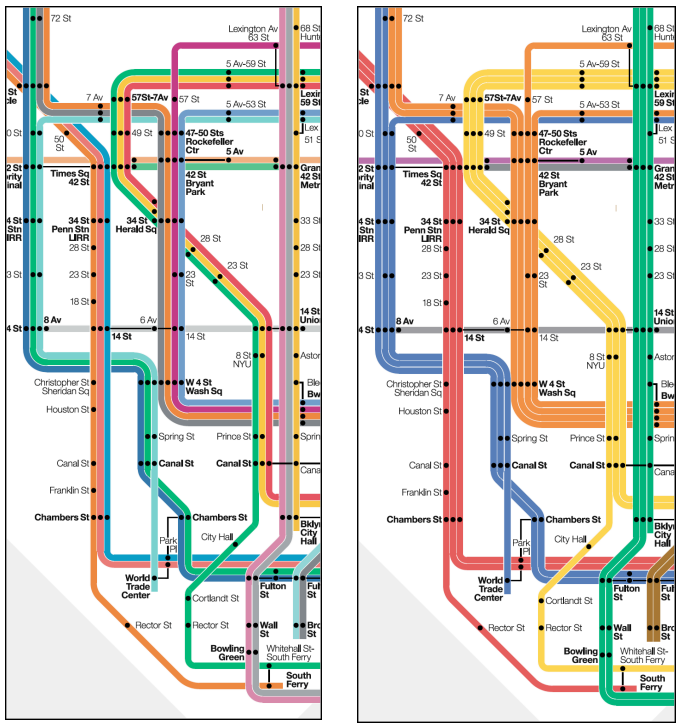


Fig. 1. Sections of maps from Lloyd, Rodgers, and Roberts [11] showing *route colour coding* (left) and *trunk colour coding* (right).

[1] See [14] for a comprehensive discussion of the nature of schematised maps of transport networks, and [13] for many examples.

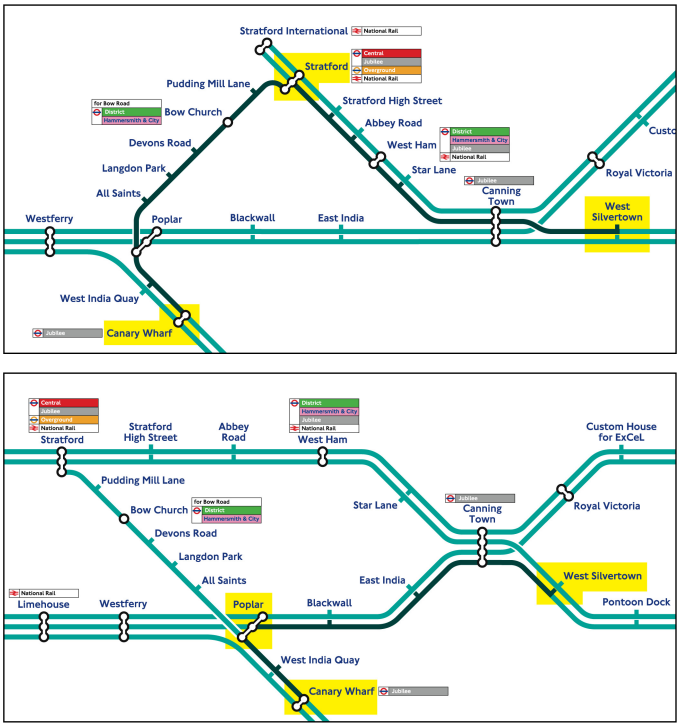


Fig. 2. Enlarged sections from the DLR maps tested by Roberts and Rose [18]. The first map (top) tended to be associated with inefficient journey suggestions, for example from Canary Wharf to West Silvertown indirectly via Stratford. The second (bottom) tended to be associated with more efficient journeys, in which subjects suggested changing at Poplar and travelling via the more direct route.

large numbers of errors. However, even in circumstances in which errors are rare, there is still potential for designs to differ in their effectiveness. Numerous studies [8, 18] have shown that line configuration can be associated with inefficient route choices so that, although such journeys are technically feasible, they are somewhat circuitous, and there are faster, more direct options available (see Fig. 2). In terms of practical utility, proportions of inefficient journeys can be conceptualised similarly to proportions of erroneous journeys, their prevalence providing another measure of usability, such that the worst-offending maps are difficult to use in the sense that optimal routes are hard to identify. In terms of psychological utility, users will generally wish to travel by optimum routes, although other factors may come into play as experience is gained in using a network, such as a desire to avoid unpleasant travel conditions on the busiest routes.

Other usability research [e.g., 15, 16, 17, 19] has compared maps whose design is *basically competent*, i.e. error rates are low, and in which they do not differ in efficiency of planned journeys. For these, *journey planning time* is taken as the prime measure of usability. Hence, given origin station A and destination station B, how long does it take to plan a route between the two? In such studies, differences in planning times between maps can be substantial, with instances of less-effective maps requiring, on average, up to twenty percent more planning time than more effective maps. With journey planning time as an index of difficulty, such that it is assumed that slower maps have general configural weaknesses that make them harder to use, it therefore seems non-controversial to categorise such designs as being less desirable. However, an element of caution is required because, compared with the other objective measures discussed, there is potential for a disconnection, between journey planning time, and the actual task of navigating an urban rail network. Using a schematic map is a complex, multi-faceted task, so we need to be sure that journey planning time does, indeed, have practical utility, rather than indexing an aspect of design that is of low relevance. In other words, suppose Map A yields a mean journey planning time of 50 seconds per journey, and Map B yields a mean of 40 seconds, with the journeys themselves identical. How likely is if that the 20% difference between maps points towards usability differences that actually matter?

II. THE USABILITY GAP

Usability testing of schematic maps is resource hungry. For example, investigating various prototype train diagrams for London's Docklands Light Railway, Roberts and Rose [18] identified designs that were prone to either errors or circuitous journeys but, for this relatively simple network, 240 people were tested individually for approximately 20 minutes each. In such circumstances, it might be tempting simply to present people with maps and ask them to select preferred designs. Indeed, in this study, people were also asked to indicate preferences, and to complete questionnaires in which designs were rated. However, neither of these methods of ascertaining subjective views was able to flag those designs particularly associated with erroneous or circuitous journeys. Not surprisingly, the people who were making these journey planning errors/inefficiencies were unaware that the maps that they were using were inducing them.

Public votes are occasionally solicited in order to choose between schematic maps, most notably in Boston in 2013 [3]. However, research has demonstrated considerable diversity of opinions on the criteria for usability which, by itself, would entail caution before evaluating prototypes in this way. For example, Roberts, Gray and Lesnik [15] found that, in rating various London underground maps for usability, of those people whose ratings were the most internally consistent, a group could be identified comprising people who placed the most weight on simplicity of line trajectories, and another group (approximately twice the size) could be identified whose members placed the most weight on the design rules (octolinear maps preferred, i.e. those utilising only horizontal, vertical, and 45º diagonal straight lines only).

More seriously, in considering journey planning times, various studies by Roberts and colleagues [15, 16, 17, 19] have found no relationship between people's opinions of maps, and their usability. For example, in certain studies [15, 16, 19] people planned journeys using two different designs. The individuals directly experienced using both, which might assist in callibrating subjective judgments, and this also yielded two mean journey planning times for each individual, one for each map type. It is therefore possible, for each person, to identify the particular map that was easiest to use on this measure. Unfortunately, determining the most effective design objectively on an individual basis does not relate to subjective ratings in any intelligible way. For example, looking at simple map preferences, we might expect individual choice to reflect the personal relative planning time advantage that one design of a pair has over the other. The faster map should be selected in preference. Table I shows not only that there is no statistical

TABLE I. Data from three usability studies, each showing that there is no significant relationship between relative map usability at an individual level, as indicated by journey planning times, and individual map preference.

Roberts & Vaeng [19]: London curvilinear compact vs. London octolinear compact	Mean planning time advantage for curvilinear map
Chose curvilinear ($N = 8$)	17.5 sec
Chose octolinear ($N = 14$)	8.1 sec
	$t_{(20)} = 1.08$, $p > .05$
Roberts, Gray & Lesnik [15, Experiment 1]: Paris curvilinear vs. Paris official octolinear	**Mean planning time advantage for curvilinear map**
Chose curvilinear ($N = 34$)	8.4 sec
Chose octolinear ($N = 38$)	9.7 sec
	$t_{(70)} = 0.37$, $p > .05$
Roberts, Newton & Canals [16]: Berlin octolinear vs. Berlin concentric circles	**Mean planning time advantage for octolinear map**
Chose octolinear ($N = 33$)	6.0 sec
Chose concentric circles ($N = 6$)	4.7 sec
	$t_{(37)} = 0.57$, $p > .05$

corroboration for this, there is not even a consistent trend in the expected direction. This is still the case when looking at more detailed measures, such as aggregated scores from rating questionnaires. In all these studies, the correlation between subjective map preference, and objective map performance, was effectively zero. Hence, large numbers of users were selecting ineffective designs and rejecting effective ones. In other words there is potential conflict owing to a *usability gap* that leads to disagreement concerning the maps that users *want to use*, and the maps that researchers believe users *ought to use*.

III. EXPLAINING AWAY THE USABILITY GAP

The lack of ability of people to identify the most effective designs as a group, or individually, is not unprecedented in the psychological literature. Indeed, in other applied settings, such as usability of weather maps [9] and human computer interaction [1] there are also tendencies to prefer less effective representations/methods and reject more effective ones. There is a large body of research on the general fallibility of human judgements in various contexts. Of relevance is the possibility that people's expectations and prejudices about map design override any actual experience of maps during usability testing. Hence, an octolinear map may evoke strong preferences because people encounter these, sometimes on a daily basis, all round the world. On the other hand, there are good reasons why it might be difficult to override people's preferences. For example, a recurring finding in the literature on expert versus novice judgements is that novices' evaluations tend to focus on superficial surface details – which for a schematic map might be whether it conforms to expectations, or its use of colours – whereas experts in a domain are better able to base judgements on underlying theoretical principles [6].

Another reason why users might have difficulty to override their expectations, and identify effective designs on the basis of journey planning times, is that this is an extremely difficult task, such that users effectively suffer from a *metacognitive deficit*. Metacognition refers to people's ability to have an awareness or understanding of their own cognitive processes, for example, the procedures that they used in order to make a decision, the factors that influenced them, and their level of performance. There is a large literature pointing towards many general and wide-ranging weaknesses that people have.

Probably the most well-known and dramatic demonstration of a metacognitive deficit is by Chabris and Simons [5]. The task was to watch a video of people playing with basketballs and count the passes made by the people dressed in white (as opposed to black). During this, an actor dressed in a black gorilla costume walks across the frame, makes a gesture to the camera, and leaves. The gorilla is almost always missed but, of particular interest here, is people's disbelief that they could have missed the gorilla, indicating limited awareness of their cognitive (in)capabilities. Also relevant, Kruger and Dunning [10] showed that in a variety of cognitive and social domains, people are poor at judging their level of performance, with the worst performers showing the biggest gap between assessed performance and actual performance. On the topic of choices, Nisbett and Wilson [12] asked people to select between various options and, subsequently, explain their choices. People generally justified these in terms of properties, such as colour and taste, but the experimenters found that, generally, choices were predictable from trivial variables such as positioning, concluding that people have limited awareness of their basic cognitive processes and, instead, construct *post-hoc* narratives to account for their choices retrospectively. Studies such as the three discussed all show that lack of insight into performance and cognition is by no means unusual.

Even worse for the user who is attempting to identify the most effective designs of schematic map on the basis of planning times, metacognitive monitoring of self-performance is a demanding task that requires cognitive resources, which are already being expended on the task itself [7]. Also, in the absence of positive information (such as a stopwatch in view) time is a notoriously difficult variable to estimate, especially as its perception depends on the difficulty of the task being performed [2]. Hence, faced with the usability gap, it is straightforward for the psychologist to identify a substantial literature showing that a zero correlation between subjective ratings and objective measures is by no means surprising. Hence it can be argued that the subjective assessments of the users are simply in error.

IV. THREE CLUES TO UNDERSTANDING THE USABILITY GAP

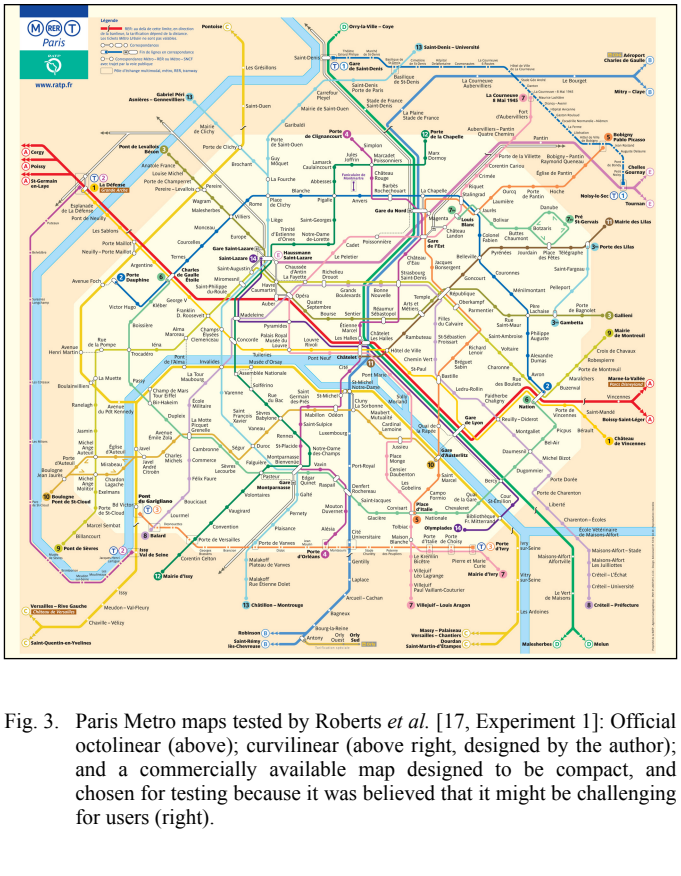
Although it is easy to explain away the usability gap, it is advisable to hesitate before taking this route for three reasons. First, from a psychological perspective, the zero correlation between objective measures and subjective ratings should, at the very least, be unsettling. Small correlations would normally be expected, but no relationship at all hints at a massive degree of orthogonality between objective measures versus the user, such that the former are actually irrelevant to daily usage and needs. This was hinted at in Section I, noting that planning times (even differences of tens of seconds between maps) were only obliquely related to actual consequences for using an urban rail network. Second, the usability gap implies a very superficial analysis of maps by users, but this is clearly not the case. For example, Roberts, Grey and Lesnik [15, Experiment 2] asked people to rate designs *separately* for usability and attractiveness. If ratings were a mere expression of at-a-glance

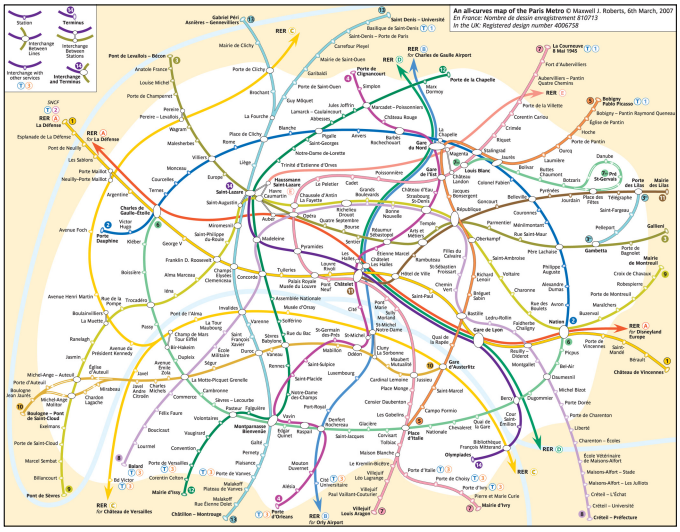
first impressions, the two should be highly correlated, but in fact a strong dissociation was identified: Multilinear maps were rated as being more usable than curvilinear maps, and curvilinear maps were rated as being more attractive than multilinear maps. This suggests a reasonably sophisticated analysis by users that warrants further attention. Third, irrespective of objectively measured usability, user-acceptance of maps is of key importance. An impeccably usable map will nonetheless fail if users reject it and seek alternatives. It would be desirable to seek some sorts of bridge between objective measures versus subjective ratings, such that common ground can be identified, making the current situation less adversarial. For these reasons, it is worth taking a closer look at the methods used to investigate journey planning times, and at some of the finer details of the studies.

1) *Users do not seem to care about journey planning times*

The obvious solution to the difficulties that people have in perceiving time is to heighten the salience of this variable and, if journey planning time is perceived as being insignificant for all practical purposes, then to heighten its importance. Several unpublished studies by the author have attempted, in various ways, to bridge the gap between journey planning times and subjective evaluations of maps. These include (i) *visible timers*, so that users can have full awareness of their planning time durations for each journey for each map; (ii) *self-estimated times*, in order to see whether, at the very least, subjective perception of journey planning time is correlated with subjective ratings of designs; (iii) *deadline tasks*, in which users are asked to imagine planning a journey while a train is pulling into a platform, and hence only limited time is available to plan a journey; and (iv) *false timers with deadlines*, in which the speed of countdown is altered systematically across maps – e.g., for Map A the twenty second countdown takes 17.5 seconds, versus 22.5 seconds for Map B, and vice-versa – in an attempt to manipulate time-outs and relative perceived failure of the designs. Unfortunately, in every case, these attempts to bridge the usability gap, by heightening the salience and importance of journey planning time, have failed. For example, for deadline tasks, even where particular designs are associated with many time outs, and hence planning failures, there is no tendency for those maps associated with more time-outs to be more adversely rated by individuals.

Of course, it is perfectly possible that the manipulations described above failed to yield correlations, between various aspects of journey planning timing and subjective ratings of map usability, for methodological or implementational reasons. Hence it cannot be ruled out that future studies might report improved methodology in which metacognitive awareness of journey planning time, and its importance, is sufficiently heightened to enable correlations between performance and user-ratings to be identified. In the interim, there seems to be good grounds to conclude that journey planning time has low psychology utility. In other words, within reason, users just do not care about planning time differences of a few seconds between maps, no matter how statistically significant. Their counter-arguments to results found using this variable might be along the lines that such planning time differences *pale into insignificance compared with the tens of minutes that might be lost if an inappropriate journey is attempted*.





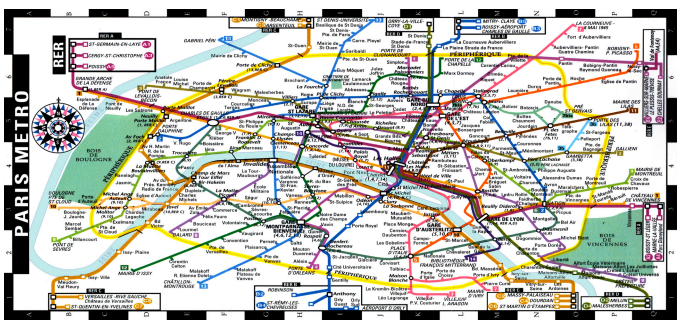


Fig. 3. Paris Metro maps tested by Roberts *et al.* [17, Experiment 1]: Official octolinear (above); curvilinear (above right, designed by the author); and a commercially available map designed to be compact, and chosen for testing because it was believed that it might be challenging for users (right).

TABLE II. Data from Roberts *et al.* [17, Experiment 1] comparing the three maps in Fig. 3 on various objective and subjective measures. There were significant differences between designs for all four of these.

	Official octolinear	Curvilinear	Commercial
Mean planning time (seconds per journey)	66.3	52.4	63.9
Mean estimated journey duration (minutes)	58.9	60.2	62.1
Mean aggregate questionnaire score (11 to 77, high = better)	56.4	56.0	43.4
Percent invalid routes	6.5	2.0	10.0

2) Users and researchers agree on the very worst maps, but the journey planning time data do not reflect this

Roberts *et al.* [17, Experiment 1] tested three maps (see Fig. 3). The official octolinear Paris Metro map, a curvilinear map designed by the author, and a third map, unofficial but commercially available, that the author felt would be particularly challenging for users: it was lacking in any clear structure, over-compact, congested and had numerous station names obstructing lines. The questionnaire ratings of the map were poor (see Table II) indicating that the users were in full agreement with the experimenter concerning this design, but the supposed difficulty of this map did not manifest itself in terms of journey planning times, where it was equal to the official octolinear map (both maps were significantly slower than the curvilinear map). Instead, the adverse performance manifested itself in terms of (i) the most invalid journeys, and (ii) estimated journey efficiency – which was calculated from a simple count of two minutes per station and ten minutes per interchange for the chosen routes). The differences are small, but suggest that people were struggling to identify efficient valid journeys using this design. Irrespective of the correct interpretation, the deficiencies of this map did not manifest themselves in terms of uniquely bad journey planning times.

3) Users seem to make valid complaints about circles maps

Roberts, Newton and Canals [16] compared two Berlin maps: one was based on concentric circles, the other was a conventional octolinear design (see Fig. 4). This is one of the few usability studies were user reactions corresponded with data: The concentric circles map was rated poorly by users, and had significantly slower journey planning times than the octolinear map. Even so, subjective measures and journey planning times were nonetheless uncorrelated (e.g., see Table I), indicating that the causes of prolonged journey planning times were different to the causes of the adverse ratings. In this study, user comments included a recurring complaint that the structure of the concentric circles map made every option look roundabout, and hence it was difficult to identify efficient journeys from amongst competing alternatives.

V. BRIDGING THE USABILITY GAP

When evaluating data, it is often useful to make a distinction between statistical significance versus practical significance. Conventional measures of task difficulty, such as proportions of invalid or inefficient journeys, can have direct practical utility, but other proxy measures of map effectiveness, such as journey planning time, may not be capturing aspects of usability that have this. Hence, a few seconds difference in journey planning time between maps would only be materially important if, for example, these fed through to differences in the selection of efficient journeys when multiple routes are available. Conversely, a few seconds journey planning time advantage for a particular map would be irrelevant if journeys identified using it tended to be inefficient.

In the real world, usability is complex and multifaceted, and care is needed in attempting to measure it. Certain concepts are easy to identify and measure, but it is also important to establish their validity. An analysis of the usefulness of journey planning time as an index of task difficulty reveals that its legitimacy might be questioned. Irrespective of the outcome, a searching analysis of such issues can only be beneficial either (i) strengthening the status of measures of performance, or else (ii) leading to their replacement by improved ones. The first outcome would support the *metacognitive deficit* explanation of the usability gap, the second might bridge the gap if replacement measures proved to be correlated with subjective assessments of schematic map usability.

Fig. 4. Sections of maps from Roberts, Newton and Canals [16]. Each has an origin/destination station pair highlighted, and two alternatives for travel between them. For the octolinear map (top), Route A (via the blue *Stadtbahn*) appears to be indirect, and Route B (via *U1*, green) would be preferred. For the concentric circles map (lower) both routes appear circuitous, and it is difficult to identify the best option.

There are two reasons why it is important to try to bridge the usability gap in schematic mapping. First, it is inherently adversarial in nature. Effectively, users are judging schematic map usability, but researchers are declaring that their assessments are misguided. As we have seen, people's self-beliefs in the efficacy of their own judgments are generally very strong, likewise their lack of interest in journey planning time as a measure of usability, hence the measure has low psychological utility. The risk of this adversarial situation is that people, perhaps including those commissioning maps, such as transport professionals, simply reject usability testing outright as ignoring the real needs of users, preferring methods that they have more faith in, such as questionnaire ratings and focus groups. Second, if attempting to bridge the usability gap results in new measures of map effectiveness that have high psychological and practical utility, then this offers the potential for improved evidence from which designs may be better optimised in the future. It also offers the possibility of more cost-effective usability testing, in which legitimate opinions of users are targeted via questionnaires and other tasks. The presence of at least some correlation between people's opinions on design and objective data will also improve the chance of such research being taken seriously by skeptical individuals.

The overall conclusion of this paper is that there is the potential need for new measures of schematic map usability to be developed. Journey efficiency discriminability is suggested as a candidate for this on the basis of a number of observations across several usability studies. People's ability to discriminate between alternative journeys may vary from map to map, and also be correlated with users' evaluations of these. Of course, evaluating designs on the basis of such a measure presents many challenges, not least because a schematic map potentially distorts spatial relationships between stations. Hence a visual estimation of relative journey efficiency between options on a map need not comply with actual reality. Should journey efficiency discriminability prove to be a useful measure of usability, this in turn would highlight the need for a fuller understanding of the effects of topographical distortion on schematic map usability, with the proviso that, for a complicated network such as London or Paris, greater topographical accuracy may be associated with more complicated line trajectories, in turn making the relative effectiveness of different routing options harder to evaluate.

References

[1] Andre, A.D., & Wickens, C.D. (1995). When users want what's not best for them. *Ergonomics in Design*, *3*, 10-14.

[2] Block, R.A., Hancock, P.A., & Zakay, D. (2010). How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica*, *134*, 330-343.

[3] Boston Globe (2013). MBTA map-making contest garners 17,000 votes. http://www.bostonglobe.com/metro/2013/09/21/map-making-contest-garnersvotes/egeCO7x7Q8rbfciGVHaSyI/story.html [accessed 20/03/2019]

[4] Bronzaft, A.L., Dobrow, S.B., & O'Hanlon, T.J. (1976). Spatial orientation in a subway system. *Environment & Behavior*, *8*, 575–594.

[5] Chabris, C., & Simons, D. (2010). *The invisible gorilla*. Crown Publishing.

[6] Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121–152.

[7] Dierckx, V., & Vandierendonck, A. (2005). Adaptive strategy application in linear reasoning. In M. J. Roberts, & E. J. Newton (Eds.), *Methods of thought: Individual differences in reasoning strategies* (pp. 107–127). Psychology Press.

[8] Guo, Z. (2011). Mind the Map! The Impact of Transit Maps on Travel Decisions in Public Transit. *Transportation Research Part A*, *45*, 625–639.

[9] Hegarty, M. (2013). Cognition, metacognition, and the design of maps. *Current Directions in Psychological Science*, *22*, 3-9.

[10] Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it. *Journal of Personality & Social Psychology*, *77*, 1121–1134.

[11] Lloyd, P.B., Rodgers, P., & Roberts, M.J. (2018, June). Metro Map Colour-Coding: Effect on Usability in Route Tracing. *In International Conference on Theory and Application of Diagrams* (pp. 411-428). Springer, Cham.

[12] Nisbett, R.E., & Wilson, T. DeC. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231-259.

[13] Ovenden, M. (2015). *Transit maps of the world* (3rd ed.). Penguin Books.

[14] Roberts, M.J. (2012). *Underground maps unravelled: Explorations in information design*. Published by the author.

[15] Roberts, M.J., Gray, H., & Lesnik, J. (2017). Preference versus performance: Investigating the dissociation between objective measures and subjective ratings of usability for schematic metro maps and intuitive theories of design. *International Journal of Human Computer Studies*, *98*, 109-128.

[16] Roberts, M.J., Newton, E.J., & Canals, M. (2016). Radi(c)al departures: Comparing conventional octolinear versus concentric circles schematic maps for the Berlin U-Bahn/S-Bahn networks using objective and subjective measures of effectiveness. *Information Design Journal*, *22*, 92-114.

[17] Roberts, M. J., Newton, E. J., Lagattolla, F. D., Hughes, S., & Hasler, M.C. (2013). Objective versus subjective measures of Paris Metro map usability: Investigating traditional octolinear versus all-curves schematic maps. *International Journal of Human Computer Studies*, *71*, 363-386.

[18] Roberts, M. J., & Rose, D. (2016). Map-induced journey-planning biases for a simple network: A Docklands Light Railway study. *Transportation Research Part A: Policy and Practice*, *94*, 446-460.

[19] Roberts, M.J., & Vaeng, I.C.N. (2016). Expectations and prejudices usurp judgements of schematic map effectiveness. In: P. Lloyd & E. Bohemia (eds.), *Proceedings of DRS2016: Design + Research + Society — Future-Focused Thinking*, *Volume 8*, pp 2343-2359.